系统生物学中的物理问题*

欧阳颀†

（北京大学物理学院介观物理国家重点实验室　北京大学理论生物学中心　生命科学联合中心　北京　100871）

摘　要　在20世纪末到21世纪初的十多年里，生命科学，特别是分子生物学发生了令世人瞩目的变化.生命科学研究飞速发展使人们相信21世纪是生命科学的世纪.与此同时，人们也越来越清楚地意识到生命科学研究的质的飞跃不可能由生物学家独立完成.数学、物理、化学、力学、信息科学在生物学研究中必将担任越来越重要的角色.文章通过介绍几个作者参与的系统生物学研究工作，探讨物理学在系统生物学中应该并能担任的角色.

关键词　系统生物学，定量生物学，非线性动力学，反向工程

The physics questions in systems biology

QUYANG Qi†

(*State Key Laboratory for Mesoscopic Physics and School of Physics*, *Center for Theoretical Biology*, *Center for Life Sciences*, *Peking University*, *Beijing* 100871, *China*)

Abstract　In the last decade between 20th and 21st century, life sciences, especially molecular biology have witnessed a dramatic change. As a result, it is generally believed that the 21 century is the century of life science. At the same time, it was realized that the leap forward in life sciences cannot be accomplished by biologists alone. Mathematics, physics, chemistry, mechanics, information science were all needed in this process. In this article, we use a few examples of our works to elaborate the physics role in life science researches.

Keywords　systems biology, quantitative biology, nonlinear dynamics, reverse engineering

1　物理学在现代生命科学中的作用

上世纪60年代，一位美国的著名理论物理学家曾经效仿当时美国总统肯尼迪的语言说过这样一句话：不要问物理学家能为生命科学家做什么，要问生命科学家能为物理学家做什么.这句底气十足的话反映了当时物理学的辉煌成就和物理学家解析一切自然规律的自信.这样的自信在当时无疑是有道理的.从现代物理学发展的历史看，只要物理学家能够掌握足够的定量实验数据，他们总能够成功地找出隐藏在这些实验数据背后的物质运动基本规律，做出正确的预言，并指导科学技术的发展.从这个角度观察，当时的物理学家普遍认为，一旦生物学家能够为物理学家提供足够的高质量定量数据，他们一定能够找到统治生命现象的普适性基本规律.

将时间推移到本世纪初，我们发现物理学家的这种对掌握自然规律的自信受到了来自生命科学的巨大挑战.就目前物理学在现代生命科学中的作用看，上面提到的那句断言正好要倒过来：不要问生命科学家能为物理学家做什么，要问物理学家能为生命科学家做什么.倒过这句话的主要依据是：在20世纪末到21世纪初的十多年里，生命科学，特别是分子生物学发生了令世人瞩目的变化.一方面，由于基因组测序、蛋白质组学的快速发展，生物学积累了大量的数据，如何挖掘出大量实验数据所蕴藏的生物基本规律已为生命科学以至于整个科学研究的焦点；另一方面，研究生物学系统的信息处理过程开始从对单一信号传导通路的定性描述转移到对复杂蛋白质与基因调控网络的定量刻画.在大量数据面前，物理学家至今还在探索这些数据背后隐藏的基本定量规律.到目前为止，这方面的工作还没有决定性的突破.生命科学从以定性描述到定量刻画的转变很可能在生命科学研究领域产生革命性的变化，而这个革命性的变化需要人们对生命物质运动的基本规律有精确的掌握，能够对生命现象做出定量的预测.这个任务显然是物理学家责无旁贷的.

*　国家自然科学基金（批准号：11074009，10721463）、国家重点基础研究计划（批准号：2009CB918500）资助项目
2011－10－06收到

†　Email：qi@pku.edu.cn

实际上,生物系统表现出的形态多样性与系统稳定性很早就引起了物理学家的关注.在科学史上用各种理论解释生命现象的尝试从来就没有停止过.用现代物理学观点揭示生命现象的早期尝试之一是薛定谔的著作《什么是生命》,虽然用现在的观点看薛定谔对生命现象的解释不很准确,但他的书的确吸引了一批物理学家的目光,引导他们用定量的手段研究生命现象,并试图找出它们中间的普遍规律.从物理学角度考察,生命系统可以被看成是一个复杂的非线性动力系统.活体细胞的分子(如蛋白质、基因等)网络控制系统根据细胞的初始条件、边界条件和环境变化,通过激活或抑制一些功能性蛋白的产生来改变自身的状态,以实现某种生物功能.随着分子生物学研究的发展,尤其是定量生物技术的发展,定量描述这类动力系统的研究已经逐渐成熟.目前国际上与生命科学交叉的一个重要的生物物理研究方向就是定量研究蛋白质相互作用及基因调控网络的拓扑结构、动力学性质、生物功能以及它们之间的关系,这构成了系统生物学研究的一个重要分支,甚至是主要分支.这方面的研究至少包含两方面的研究内容:第一方面,从生物调控网络的拓扑结构出发,在此基础上建立调控网络的动力学模型,从而解释生物控制系统的动力学性质,并对系统的功能做出预测;第二方面,从生物调控网络所要执行的功能与动力学要求(如稳定性)出发,推断调控网络的拓扑性质,对控制网络的结构做出预测.本文通过介绍北京大学理论生物中心在这两个领域所做的一些初步工作,探讨物理学在系统生物学中应该并能够担任的角色.

2 生物调控网络动力学研究

生物调控网络动力学研究的核心是通过对生物的网络控制系统进行非线性动力学分析,从非线性物理及动力学角度对该系统在分子水平上进行系统的、定量的理论与实验研究.在研究中力图发现支配生物调控网络的基本动力学规律,并总结出一套适用于系统生物学研究的非线性动力学研究工具.这个方向的研究成果可能成为其他研究领域的理论基础.例如,哺乳细胞的癌变机理一直是困扰癌症研究的问题.一种癌细胞的基因突变谱常常看起来毫不相关.但是,最近的几项实验研究表明,癌症的基因突变有可能从生物分子控制网络中找到线索[1].从非线性动力学角度看,细胞癌变可能被看成是相应动力系统的非线性动力学分岔现象.而分岔的条件可能对应于控制细胞行为的控制常数改变.这里介绍北京大学理论生物学中心的一个研究结果[2],酵母蛋白质网络的动力学性质的研究,说明生物系统中的非线性动力学特征.

芽殖酵母(budding yeast saccharomyces cerevisiae)是生物学研究中广泛应用的单细胞真核模式生物,酵母在细胞周期调控的研究中有着极其重要的作用.1996 年作为第一个真核生物,芽殖酵母的全基因组测序工作完成并公布.近年来,芽殖酵母的蛋白质相互作用的数据迅速增加,这些蛋白质－蛋白质相互作用网络的数据和相关的生物学研究进展,为进一步全面系统地研究蛋白质网络的性质提供了可能.一些物理学家对芽殖酵母蛋白质网络的拓扑性质进行了研究,得出了网络连线数随网络节点成幂率分布的结论[3],另一些研究人员对基因调控网络中的基本调控单元进行了研究,试图找到生物系统中蛋白质网络的基本构成单元[4].相对于较为稳定的基因组,蛋白质网络组成的蛋白质通过对不同的环境信号和不同蛋白质的状态不断变化产生反应,即通过动力学过程完成生物学功能.所以,蛋白质网络动力学的研究成为生物学家和生物物理学家共同关心的重要问题.

芽殖酵母具有简单的生命周期,能够以单倍体和双倍体形式存在.在营养丰富的条件下,单倍体和双倍体的芽殖酵母细胞都能够以正常的细胞分裂周期进行繁殖.CDK(cyclin－dependent kinase,CDK)蛋白激酶的基因表达和活性调控了整个细胞周期过程.在营养缺乏条件下,触发孢子形成信号(sporulation signal),双倍体细胞能够通过减数分裂产生孢子,形成单倍体细胞来适应恶劣的外界条件,减数分裂主要是由 Ime1 蛋白的表达和活化来调控的.当营养丰富时,受到信息素(pheromone)的刺激,2 个单倍体细胞将融合成为 1 个新的双倍体细胞.

生物学中研究得最为清楚的是细胞周期(cell-cycle)调控网络.该网络是基于以前的动力学模型,并通过大量的文献调研和对蛋白质数据库(http://mips.gsf.de/)的分析建立起来的.简化的细胞周期网络如图 1 所示.

该网络中的蛋白质可分为以下 3 大类:第 1 类为 Cyclin/Cdc28 复合物,包括 Cln3/Cdc28 复合物(图中简写为 Cln3),Cln1/Cdc28 与 Cln2/Cdc28 复合物(简写为 Cln1,2),Clb5/Cdc28 与 Clb6/Cdc28 复合物(简写为 Clb5,6),Clb1/Cdc28 与 Clb2/Cdc28 复合物(简写为 Clb1,2);第 2 类为转录因子,包括 MBF,SBF,

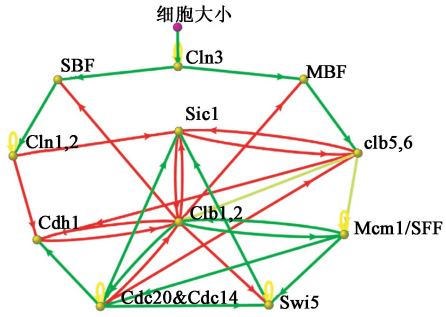

图 1　芽殖酵母细胞周期调控网络

Mcm1/SFF 和 Swi5；第 3 类为 Cyclin/Cdc28 复合物的抑制蛋白与降解蛋白，包括 Sic1，Cdh1，Cdc20/APC. 网络中绿色箭头表示正相互作用（激发或活化），红色箭头表示负相互作用（抑制或去活化），黄色箭头表示蛋白质的自降解作用（见《物理》网刊彩图，下同）. 细胞周期过程从定性的生物学角度可以简述如下：在营养丰富的条件下，当双倍体或单倍体的酵母细胞长得足够大时，Cln3/Cdc28 蛋白复合物将被活化，促使细胞进入“激发的”G_1 态，这时细胞的 Sic1 浓度较高，Hct1 处于活化状态. 活化的 Cln3/Cdc28 复合物将活化转录因子 MBF 和 SBF，活化的 MBF 和 SBF 与 DNA 结合后，转录相应的 mRNA，然后翻译形成 Cln1，Cln2，Clb5，Clb6 蛋白，上述蛋白抑制了 Sic1 和 Hct1 的作用，并控制着 G_1 后期基因的表达. 在 S 期（S phase），细胞复制自己的 DNA. 通过 G_2 期，Clb1 和 Clb2 活化，细胞进入有丝分裂期（M phase）. 有丝分裂使得复制的 DNA 等量地分配到细胞相对的两极，然后一个细胞分裂产生 2 个子细胞，在该过程中，Cdc20/APC 和 Swi5 被活化，导致 Sic1 浓度升高，Hct1 活化，并对 cyclin/Cdc28 复合物产生抑制作用. 最后，细胞又回到细胞周期的静息 G_1 态，即 G_1 基态，等待下一次分裂信号. 总的来说，细胞周期过程起始于“激发”G_1 态，使得Cln3/Cdc28复合物处于活化状态，通过一系列细胞周期过程，最后回到 Cln3/Cdc28 复合物未活化的 G_1 基态. 以上的定性描述是分子生物学家经过几十年努力得出的综合结果. 物理学家面临的问题是，怎样用动力学的观点研究这个网络系统的基本动力学性质.

为了研究蛋白质网络的动力学性质，最简单的办法是选择以下简单的离散动力学模型：每类蛋白质只有两种状态，0 与 1，分别表示该蛋白质处于活化与未活化状态. 下一个时刻蛋白质的状态是由当前时刻的蛋白质状态按照以下规则决定的：

$$S_i(t+1)=\begin{cases}1:\sum_j a_{ij}S_j(t)>0 & ,\\ 0:\sum_j a_{ij}S_j(t)<0 & ,\\ S_i(t):\sum_j a_{ij}S_j(t)=0 & ,\end{cases}\qquad(1)$$

其中 a_{ij} 是第 j 类蛋白质对第 i 类蛋白质的作用系数. 模型中的时间步长是逻辑步长，而非实际意义上的时间. 选择 a_{ij} 取值为 1 与 $-\gamma$，分别表示正相互作用（绿色箭头）和负相互作用（红色箭头）. $\gamma=1$ 为等权模型，$\gamma\gg1$ 为强抑制模型，后者更加接近于生物系统. 自降解作用（黄色箭头）具有时间延迟的性质：一个具有自降解作用的蛋白质，若在 t 时刻被活化（$S_i(t)=1$），而且在 $t+1$ 到 $t=t+t_d$ 时间内一直没有其他的正负输入，那么它将在 $t=t+t_d$ 时刻降解（$S_i(t+t_d)=0$）. 在模型中，选择 $t_d=1$. 采用简单的离散动力学模型的优势在于能够分析网络动力学状态全空间的性质，从而得到网络的全局动力学性质.

现在利用以上的离散动力学模型研究细胞周期调控网络随“时间”的变化. 首先把激发 G_1 态作为初始态（Cln3，Sic1 和 Cdh1 的状态为 1，其余蛋白质的状态为 0），计算的结果表明，系统经过 13 个逻辑步长逐步演化到 G_1 基态（Sic1 和 Cdh1 的状态为 1，其余蛋白质的状态为 0），见表 1.

表 1　芽殖酵母细胞周期演化表

过程步长	Cln3	MBF	SBF	Cln2	Cdh1	Swi5	Cdc20&Cdc14	Clb5	Sic1	Clb2	Mcm1/SFF	相位
1	1	0	0	0	1	0	0	0	1	0	0	起始状态
2	0	1	1	0	1	0	0	0	1	0	0	
3	0	1	1	1	1	0	0	0	1	0		G1
4	0	1	1	1	0	0	0	0	0	0	0	
5	0	1	1	1	0	0	0	1	0	0	0	S
6	0	1	1	1	0	0	0	1	0	1	1	G_2
7	0	0	0	1	0	0	1	1	0	1	1	
8	0	0	0	0	0	1	1	0	0	1	1	M
9	0	0	0	0	0	1	1	0	1	1	1	
10	0	0	0	0	0	1	1	0	1	0	1	
12	0	0	0	0	1	1	0	0	1	0	0	G_1
13	0	0	0	0	1	0	0	0	1	0	0	终止状态 G_1

蛋白质状态的时间演化过程与生物学实验观察相符合，这说明表 1 描述的控制细胞周期的蛋白质作用网络基本抓住了系统动力学的关键.

网络动力学研究的第一个任务是了解它的动力学吸引子. 遍历所有可能的 2048 个初始态(包括 11 个结点网络，不包括 Cell Size 信号)，该蛋白质网络最后演化到 7 个稳定的状态，其中 1764 个初始态(约 86%)演化到静息的 G_1 态，即细胞周期的生物学稳定态. 这说明细胞周期的静息 G_1 态是一个全局吸引子，而且是唯一的全局吸引子. 为了验证是否所有网络都具有类似的性质，在相同演化规则下，可以研究具有相同结点数目和相同连接数目的随机网络的吸引域分布. 随机网络的动力学性质的研究表明，随机网络平均来说具有更多的吸引子，而且出现类似细胞周期网络全局吸引子的概率极小. 因而细胞周期网络的特殊结构使得其生物系统能够具有很好的全局动力学稳定性.

那么，11 个结点的细胞周期网络中的所有初始状态是怎样一步一步地演化到最后的吸引点呢？图 2 中给出了所有的 2048 个初始状态的演化路径，图中最粗的路径为生物学路径——细胞周期路径，最大的节点为 G_1 基态，图中每个点的大小和每条线的宽度正比于 $\ln(2+m)$，m 为蛋白质初态经过该点(边)的数目. 我们发现，细胞周期路径是一条一维稳定流形：大部分的网络初始状态首先被吸引到细胞周期路径上来，然后沿着细胞周期路径逐步到达稳定态——G_1 基态. 这意味着不仅 G_1 基态是一个全局性的稳定点，而且从 G_1 激发态到 G_1 基态的细胞周期路径同样是一个全局性的稳定的动力学路径. 蛋白质网络通过长期的进化，其动力学性质具有双重稳定性. 更加深入的研究发现，在裂殖酵母(fission yeast)和蛙卵细胞(frog egg)的细胞周期网络中也有类似的动力学性质.

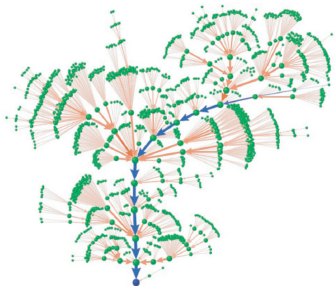

图 2　牙殖酵母细胞周期状态在相空间的演化路径

生物物理的实验表明，细胞体内的蛋白质相互作用与蛋白质－DNA 相互作用能量一般都在 3—5kT 之间. 这表明细胞内的分子动力学过程一直在经受环境的猛烈“轰炸”. 生命物质在经历长时间演化后找到了一些特别的相互作用结构(调控网络)，使得它们的演化轨道和静息态都是全局稳定的. 这就给予了生物调控网络系统最大的健壮性(robustness). 全局稳定性与路径稳定性都是典型的非线性现象，是非线性物理的研究领域. 但是坦白地说，非线性动力学现有的分析工具在大部分情况下只能分析系统在相空间中的局部行为，对系统全局的动力学行为的研究还刚刚开始. 在这种状况下，生物调控网络系统表现出的动力学全局稳定性的原因还没有确切的答案. 一个可能的解释是，其动力学轨迹附近存在一些鞍-节分叉点的残存流形，这些残存流形表现出的“鬼魂效应”(ghost effect)，保证了网络动力学轨道的全局稳定性. 这个解释还需要系统生物学实验的证实.

3　生物调控网络的逆向工程

系统生物学研究的另一个方向是从生物调控网络所要执行的功能与动力学要求出发，推断调控网络的拓扑性质，对控制网络的结构做出预测. 生物分子间的相互作用关系与生物功能及表型之间有很强的联系. 生物分子间的相互作用构成了生物相互作用网络，确定相互作用网络是系统生物学的一个重要任务. 通过实验确定分子间的相互作用需要大量的时间和资金，例如，确定抑癌基因 P53 和 MDM2 之间的负反馈相互作用用了 10 年的时间. 伴随着基因芯片技术的发展，理论学家提出了大量的理论方法通过高通量的实验数据去重建网络，这就是所谓的生物调控网络的逆向工程问题.

生物调控网络的逆向工程工作的基本思路是：通过基因芯片表达数据，可以得到所研究的网络节点的 mRNA 表达的时序图，经过分析可以得到对应于动力学中的一条动力学路径(如表 1 和图 2 中的蓝线所示). 由这条动力学路径，根据不同的目的，可以建立不同的模型来重建和分析网络的拓扑结构，即节点间的相互作用关系. 目前主要的分析手段可以分为基于关系[5]、基于信息论[6]、基于动力学[7]和基于贝叶斯网络[8]的分析方法. 这里介绍基础动力学的分析方法，即根据生物网络相互作用的动力学形式((1)式)与系统的动力学轨迹(图 2)，推断生物网络的拓扑结构(图 1).

这方面的开创性工作是美国乔治·华盛顿大学(George Washington University)的曾晨教授所做出的[9]. 记网络中第 i 个节点在 t 时刻的状态为 $S_i(t)$，在上述的布尔模型中，其取值为 0 或 1. 如果

网络中存在 i 节点到 j 节点的抑制作用线，记 $r_{ij}=$ TRUE(真)，否则 $r_{ij}=$FALSE(伪). 同样，如果网络中存在 i 节点到 j 节点的促进作用线，记 $g_{ij}=$ TURE，否则 $g_{ij}=$FALSE. 可以推断，对于(1)式中的强抑制模型($\gamma\gg1$)，系统动力学轨迹与网络相互作用关系应该满足下面的逻辑关系：

$$\begin{cases} S_i(t+1)=\left(\sum_{j\neq i}(S_j(t)\cdot g_{ji})+S_i(t)\cdot\bar{r}_{ii}+\overline{(S_i(t)}\cdot g_{ii}\right) \\ \qquad\cdot\prod_{j\neq i}\overline{(S_j(t)\cdot r_{ji})} \\ r_{ji}\cdot g_{ji}=0 \end{cases} \tag{2}$$

这里“+”和“∑”代表逻辑算符“OR”，“·”和“∏”代表逻辑算符“AND”，上横杠表示逻辑算符“NOT”，$j\neq i$ 为从 1 到网络节点总数 N，$S_i(t)$的值 0 与 1 分别代表逻辑式的 FALSE 和 TRUE. 系统在相空间轨迹上的所有状态(图 2 中的蓝线)在任何时间必须满足(2)式，这就是系统动力学行为对网络拓扑结构的限制. 经过一定的逻辑代数处理，这些逻辑限制可以转化为容易理解的联合正则形式(conjunctive normal form). 下面以表 1 中 Clb2 的演化路径为例说明这个推断过程. 从第一步到第二步的演化路径对网络结构的要求按(2)式表示为

$$r_{\mathrm{cln3-clb2}}+r_{\mathrm{cdh1-clb2}}+r_{\mathrm{sic1-clb2}}+\bar{g}_{\mathrm{cln3-clb2}}\cdot\bar{g}_{\mathrm{cdh1-clb2}}\cdot\bar{g}_{\mathrm{sic1-cbl2}}=\mathrm{TRUE}\ ,$$

下一步的限制为

$$r_{\mathrm{MBF-clb2}}+r_{\mathrm{SBF-clb2}}+r_{\mathrm{cdh1-clb2}}+r_{\mathrm{sic1-clb2}}+\bar{g}_{\mathrm{MBF-clb2}}\cdot\bar{g}_{\mathrm{SBF-clb2}}\cdot\bar{g}_{\mathrm{cdh1-clb2}}\cdot\bar{g}_{\mathrm{sic1-cbl2}}=\mathrm{TRUE}\ .$$

表 1 中有 12 步演化，根据(2)式对 cbl2 节点可以写出 12 个类似上面的逻辑关系式. 将这 12 个逻辑关系式用与(AND)算符连接，并将其转化成联合正则形式，得到

$$\bar{r}_{\mathrm{MBF-clb2}}\cdot\bar{g}_{\mathrm{MBF-clb2}}\cdot\bar{r}_{\mathrm{SBF-clb2}}\cdot\bar{g}_{\mathrm{SBF-clb2}}\cdot\bar{r}_{\mathrm{cln2-clb2}}\cdot\bar{g}_{\mathrm{cln2-clb2}}\cdot\bar{r}_{\mathrm{swi5-clb2}}\cdot\bar{r}_{\mathrm{clb5-clb2}}\cdot\bar{r}_{\mathrm{Mcm1-clb2}}\cdot r_{\mathrm{sic1-clb2}}\cdot\bar{g}_{\mathrm{clb5-clb2}}\cdot(\bar{r}_{\mathrm{clb2-clb2}}+g_{\mathrm{sic1-clb2}}+g_{\mathrm{Cdc20-clb2}}+g_{\mathrm{Mcm1-clb2}})=\mathrm{TRUE}\ ,$$

在这个表达式中，有逻辑乘法相隔的变量表示网络必须存在对应的调控关系；如果在此变量上有一横杠，则表示此调控关系在网络中必定不存在. 括号中的调控关系为“或”关系，表示括号内至少有一个调控关系成立. 因此，上式表示：节点 sic1 一定对节点 clb2 存在抑制关系；节点 clb5 一定对节点 clb2 存在促进关系；节点 MBF，SBF，cln2 对节点 clb2 一定不存在任何关系；节点 swi5，clb5，Mcm1 对节点 clb2 一定不存在抑制关系.

通过对其他节点的动力学轨迹做出同样的逻辑分析，可以从牙殖酵母细胞周期的动力学轨迹(图 2 中的蓝线)推断出如图 3 所示的网络相互作用关系. 其中实线表示必然存在的相互作用，虚线表示必然不存在的相互作用. 与图 1 相比，可以看到，网络中的 33 条相互作用线有 9 条被完全确定，另外有 20 条被完全排除，这说明网络动力学行为对其拓扑结构有很强的限制.

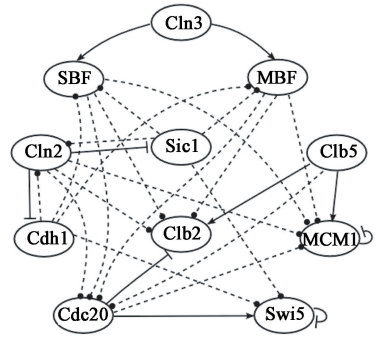

图 3　根据牙殖酵母细胞周期动力学轨迹推断出的网络相互作用关系图

一般来讲，从一条动力学轨迹得到的信息，不足以完全确定其网络节点的相互作用关系. 例如，对于如图 1 所示的 11 个节点的生物控制网络，如果考虑布尔动力学模型，则可能的网络总数为 $3^{11\times11}$ 个，即 5×10^{57} 个. 经过牙殖酵母细胞周期轨道(图 2 中的蓝线)的限制，网络总数减少为 4×10^{30} 个. 从原则上讲，假设生物调控网络没有功能对称性，图 2 所示的所有流形能够唯一确定其网络的拓扑结构. 但是实验上得到如图 2 所示的流形图需要遍历所有的初始条件，即 $2^{11}=2048$ 种不同的实验，这显然是不现实的. 能否有更好的实验设计路径，使得人们做最少的实验得到最大的网络信息，这是生物学家对物理学家提出的问题.

显然，从不同的初始条件出发会增加一条新的动力学路径. 每增加一条新的动力学路径，就会对网络拓扑结构做出新的限制. 因而问题转化为如何选择新的实验初始条件，从而用最少的实验完全确定网络的拓扑结构. 北京大学理论生物学中心在这方面做了一些系统性工作[10]，在这里做一个简单介绍.

最直观的选择办法是最大距离法. 其思想是新的动力学轨迹应该离已知的动力学轨迹越远越好，这有利于得到更多的信息. 出于这样的考虑，可以用如下方式选择新的实验条件：首先定义两个状态之间的距离 $d_{ij}=\sum_k(S_i^k-S_j^k)^2$，这里 d_{ij} 表示状态 i 与状态 j 之间的距离，S_i^k 代表蛋白质 k 在网络状态 S_i 中的值. 进而可以定义网络任意初始状态 S_i 到已知轨迹 Ξ 的最短距离：$D_i=\mathrm{Min}_{j\in\Xi}(d_{ij})$. 如果认为此距离越长，得到的

信息越多，显然应该选择最长的 D_i，即 $D=\mathrm{Max}_i(D_i)$.

另一种想到的方法是轨迹熵方法. 其基本思想是：从新的初始条件出发得到的新的动力学轨迹应该最大地减少对网络结构的不确定性. 对于牙殖酵母细胞周期调控网络，生物学路径（见表 1）将可能的网络数从 5×10^{57} 个减少到 4×10^{30} 个. 因而对于每一个初始条件遍历所有可能的网络可以产生 4×10^{30} 条动力学轨迹. 这些轨迹可能是相同的. 假设这些网络可以产生 k 条不同的动力学轨迹，其中 N_j 个网络在一个初始条件下可以产生同一条轨迹 Ξ_j. 对于此初始条件可以这样定义轨迹熵：$E=-\sum_i^k p_i\log p_i$，其中 $p_j=N_j/\sum_{j=1}^k N_j$. 显然，一个初始条件对应的轨迹熵越大，选择此初始条件做新的实验得到的网络信息就越多. 在生物网络的特性没有任何信息的条件下（即假设所有可能的网络都是等权的），这显然是最优策略. 然而在操作中，这种方法需要进行海量的计算，这是个典型的 NP 问题. 这就需要选择不同的方法估计轨迹熵.

第一种方法是分步法. 记 E_i^n 为初始条件 i 下系统发展 n 步所产生的轨迹熵. E_i^1 很容易计算. 如果假设动力学轨迹对任何网络的限制作用都是相同的，就可以从 E_i^1 经过递推计算得到 E_i^n. 具体计算见文献[10]. 这种算法的问题是其假设在许多情况下并不成立；另一种算法是采样法，即在可能的网络中进行随机采样来计算轨迹熵. 我们的计算表明，虽然网络数目巨大（4×10^{30}），对于牙殖酵母细胞周期调控网络，随机选取可能的 10000 个网络，就可以比较精确地得到任何初始条件的轨迹熵近似值.

图 4 给出了各种选择初始条件的方法的效率比较. 可以从这个结果看到如下几个特点：(1)理性选择实验初始条件可以大大减少试验次数. 对于牙殖酵母细胞周期调控网络，8 到 10 次实验就可能得到网络结构的绝大部分信息；(2)即使遍历所有的初始条件，也不能把可能的网络总数缩减为 1. 这是因为一般来讲，生物调控网络存在功能对称性，单从系统的动力学行为出发一般是不能完全断定网络的拓扑结构的；(3)在所有可能的网络都是等权的这一假设下，用采样法计算轨迹熵达到的效果最好，随机法得到的效果最差；(4)轨迹熵算法与最佳选择仍有距离，这是因为“所有可能的网络都是等权的”这一假设在实际中并不成立. 从上一节的讨论知道，生物调控网络总是选择稳健性（robustness）高的网络拓扑结构. 如何把这一性质考虑生物调控网络的逆向工程研究中，是北京大学理论生物学中心的研究任务之一.

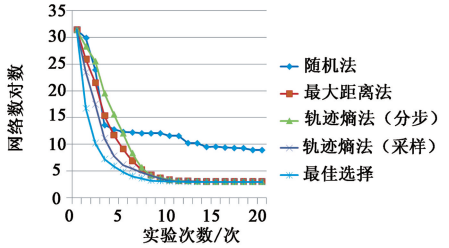

图 4　各种选择初始条件的方法的效率比较

4　展望

生命系统显然属于复杂系统. 生物调控网络在各个层次都显示出了它的多样性，即非线性与强关联特性. 从上面举出的两个例子中不难发现，现在定量研究生命现象的方法很难应用于更加复杂的系统，这在物理上被称为 scale up 问题. 这表明单从动力学角度定量研究生命系统有很大的局限性. 不管建立何种动力学模型（如布尔模型、常微分方程或偏微分方程、随机过程方程等），都会遇到计算“爆炸”问题. 同时生物系统的普遍规律很可能会在研究方程的细节时被忽略或丢失. 解决此问题的出路之一是建立生命系统的系综模型. 回顾物理学的发展，在热力学与统计物理诞生之前，在理论上可以用牛顿定律计算粒子的运动轨迹，从而得出粒子的能量与动量分布，但这种“笨”方法面对 10^{23} 个分子组成的系统显然是无法操作的. 只有在统计物理的系综理论出现后才抓住了本质，从根本上解决了这个问题. 目前物理学家在定量研究生命系统时面临相似的问题，因而建立生命系统的系综理论可能是解决问题的关键所在. 研究蛋白质相互作用及基因调控网络的拓扑结构、动力学性质、生物功能以及它们之间的关系，是系统生物学研究的内容，它是建立生命系统系综理论的准备工作，只有掌握了生命系统运动的微观规律，才可能发展出适合描述生命系统的统计理论. 物理学家在这方面还有很长的路要走.

参考文献

[1] Hayden E C. Nature, 2008, 455: 148
[2] Li F T, Long T, Lu Y *et al*. PNAS, 2004, 101: 4781
[3] Albert R, Jeong H, Barabasi A L. Nature, 2000, 406: 378
[4] Ihmels J, Friedlander G, Bergmann S *et al*. Nature Genetics, 2002, 31: 370
[5] Rice J J, Tu Y, Stolovitzky G. Bioinformatics, 2005, 21: 765
[6] Faith J J *et al*. PLoS Biol., 2007, 5: e8
[7] Gardner T S *et al*. Science, 2003, 301: 102
[8] Friedman N. Science, 2004, 303: 799
[9] Wang G *et al*. Proc. Natl. Acad. Sci. USA, 2010, 107: 10478
[10] Zhang X M, Wu Y L, Ouyang Q. PLoS Comp. Biol., 2011, under review